New York University Tandon School of Engineering
Computer Science and Engineering

CS-UY 6923: Midterm Exam.

Prof. Gustavo Sandoval

Directions

- Show all of your work to receive full (and partial) credit.
- If more space is required, use extra sheets of paper, marked with your name and the problem number.

1. Always, Sometimes, Never. (12pts – 3pts each)

Indicate whether each of the following statements is ALWAYS true, SOMETIMES true, or NEVER true. Provide a very short justification or example to explain your choice.

(a) The empirical risk of a model is lower than the population risk.

ALWAYS SOMETIMES NEVER

(b) When minimizing a loss function $L(\boldsymbol{\beta})$ using gradient descent, decreasing the learning rate η slows down how quickly we converge to a local minimum.

ALWAYS SOMETIMES NEVER

(c) For two random events X and Y, $\Pr(X \mid Y) \cdot \Pr(Y) = \Pr(Y \mid X) \cdot \Pr(X)$.

ALWAYS SOMETIMES NEVER

(d) Consider a multiple linear regression problem where each input data point has the form $\mathbf{x} = [x_1, x_2]$. Transform the predictor variables by adding quadratic terms, so each new data example has the form $\phi(\mathbf{x}) = ([x_1, x_2, x_1^2, x_2^2, x_1 x_2], y)$. Let L^* be the minimum training loss for the original problem and let L^*_{trans} be the minimum training loss for the transformed problem. Is $L^*_{trans} \leq L^*$?

ALWAYS SOMETIMES NEVER

2. Model Diagnosis Short Answer (8pts)

You are trying to solve a prediction problem using multivariate linear regression with ℓ_2 loss. You split the data set into a train set (80%) and test set (20%), then train the model on the train set to obtain a parameter vector $\boldsymbol{\beta}$. Using $\boldsymbol{\beta}$, you evaluate the average squared loss of the model on the train and test set, separately.

For each of the following scenarios, circle all answers that apply and provide a short justification.

(a) (4pts) The average squared loss on the train set is 1.5 and the average squared loss on the test set is 12.6. **Which of the following techniques is likely to improve your average test loss?**

REGULARIZATION FEATURE SELECTION FEATURE TRANSFORM DATA SCALING

(b) (4pts) The average squared loss on the train set is 10.2 and the average squared loss on the test set is 9.9. **Which of the following techniques is likely to improve your average test loss?**

REGULARIZATION FEATURE SELECTION FEATURE TRANSFORM DATA SCALING

3. Convexity of Regularized Regression (6pts)

In the lecture notes, we proved that the least squares regression loss $L(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$ is convex. It is also possible to show that the ℓ_2 and ℓ_1 penalty functions, $g(\boldsymbol{\beta}) = c\|\boldsymbol{\beta}\|_2^2$ and $g(\boldsymbol{\beta}) = c\|\boldsymbol{\beta}\|_1$, are convex. In this problem you will prove a statement which implies that, when these penalty functions are added to the standard regression loss, the resulting regularized loss is also convex.

(a) (6pts) Let $f(\boldsymbol{\beta})$ and $g(\boldsymbol{\beta})$ be any two convex functions. Let $h(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})$. Show that the function h is also convex.

4. Bayesian Crab Classification (10pts)

A biologist is collecting specimens from two species of crabs, S_0 and S_1. These species live in the same habitat and look similar to the human eye. To accelerate crab sorting by species, the biologist wants to develop a simple classification rule based on body measurements. She observes that the ratio of *forehead breadth* to overall *body length* differs between crabs in species S_0 and S_1. The biologist proposes to measure this ratio (denoted by R) and use it as a single predictor variable for classification.

The biologist assumes that the crab data comes from a "mixture of Gaussians" probabilistic model. In particular, she assumes that for each species, R follows a normal (Gaussian) probability distribution, with different parameters for each species. The biologist makes the following concrete observations:

- 35% of all crabs collected belong to S_0 and the remaining 65% belong to S_1.
- For crabs in S_0, the average value of R is $\mu_0 = .5$. For crabs in S_1, the average value of R is $\mu_1 = .4$.
- For both species, the standard deviation of R is $\sigma = .1$.

(a) (7pts) Suppose we collect a new crab with forehead breadth to body length ratio R_{new}. The biologist would like to assign this crab to S_0 or S_1 using the maximum a posterior (MAP) classification rule for her probabilistic model. Denote this rule by $f : \mathbb{R} \to \{S_0, S_1\}$. The rule takes as input the ratio R_{new} and outputs S_0 or S_1.

Write down all mathematical expressions that would need to be evaluated to compute f for a given input R_{new}. Your expressions do not need to be simplified, but they should not involve unknown variables besides R_{new}. **Hint:** Use Bayes rule.

(b) (3pts) The biologist collects a new crab with $R_{new} = .45$. Using the MAP rule derived above, will this crab be classified into species S_0 or S_1? **Hint:** This problem can be solved without a calculator.

5. Loss Minimization. (10pts)

For data with one predictor and one target: $(x_1, y_1), \ldots, (x_n, y_n)$, consider a linear regression model:

$$f_{\beta_0,\beta_1}(x) = \beta_0 + \beta_1 x$$

with *exponential loss*:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} e^{(y_i - f_{\beta_0,\beta_1}(x_i))^2}$$

(a) (5pts) Write down an expression for the gradient of the loss L. You do not need to simplify or put things in matrix form.

(b) (2pts) Name two algorithms/methods which could be used to minimize or approximately minimize L.

(c) (3pts) In general, is this exponential loss more or less robust to outliers when compared to ℓ_2 loss? How about when compared to ℓ_∞ loss?

6. Naive Bayes. (10pts)

Consider the following spam detection dataset.

1. (doc1) machine learning good (C1)
2. (doc2) machine good (C1)
3. (doc3) learning good (C1)
4. (doc4) evil learning good (C2)
5. (doc5) machine learning evil (C2)

The dataset consists of 5 documents, each of which is labeled as either "C1" or "C2". The class of each document is indicated in parentheses. Use the Naive Bayes classifier to classify the following document as either "C1" or "C2". New document: **"good evil machine"**. Use add1 smoothing. Show your work.

7. Decision Tree Learning. (10pts)

Given the following table of data,

(a) (6pts) Construct a decision tree that classifies the data. Use Information Gain as the splitting criterion and show your work.

(b) (4pts) What is your prediction for D15?

Day	Outlook	Humidity	Wind	Play
D1	sunny	high	weak	no
D2	sunny	high	strong	no
D3	overcast	high	weak	yes
D4	rain	high	weak	yes
D5	rain	normal	weak	yes
D6	rain	normal	strong	no
D7	overcast	normal	strong	yes
D8	sunny	high	weak	no
D9	sunny	normal	weak	yes
D10	rain	normal	weak	yes
D11	sunny	normal	strong	yes
D12	overcast	high	strong	yes
D13	overcast	normal	weak	yes
D14	rain	high	strong	No
D15	rain	high	weak	???

8. Reporting. (10pts)

Consider the following table:

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0

(a) (2pts) What is the accuracy of the classifier?

(b) (2pts) What is the precision of the classifier?

(c) (2pts) What is the recall of the classifier?

(d) (2pts) What is the F1 score of the classifier?

(e) (2pts) What is the confusion matrix of the classifier?

9. Practicum. (10pts)

Write a python function: `K_Fold_Cross_Validation(X, y, K, model)` that takes as input a dataset X and labels y, a number of folds K, and a model *model*. The function should return the average accuracy of the model on the dataset.

You may use the following functions:

(a) `model.fit(X, y)`. Fits the training data with model order p: $\hat{B} = fit(X, y, p)$.

(b) `model.predict(X)`. Predicts the values on the test data.

(c) `model.score(X, y)`. Returns the accuracy of the model on the test data.